Database management systems (DBMSs) are an important component of any data-intensive application. But tuning a DBMS to perform well is a notoriously difficult task because they have hundreds of configuration knobs that control aspects of their runtime behavior, such as cache sizes and how frequently data is flushed to disk. Getting the right configuration for these knobs is hard because they are not standardized (i.e., sets of knobs for different DBMSs vary), not independent (i.e., changing one knob may alter the effects of others), and not universal (i.e., the optimal configuration depends on the target workload and hardware). Furthermore, as databases grow in both size and complexity, optimizing a DBMS to meet the needs of new applications has surpassed the abilities of even the best human experts. Recent studies using machine learning techniques to automatically configure a DBMS's knobs have shown that such techniques are able to produce high-quality configurations, however, they need a large amount of training data to achieve good results. Collecting this data is costly and time-consuming. In this thesis, we seek to address the challenge of developing effective yet practical techniques for the automatic configuration of DBMSs using machine learning. We show that leveraging knowledge gained from previous tuning efforts to assist in the tuning of others can significantly reduce the amount of time and resources needed to tune a DBMS for a new application.

Thesis Committee:
Andrew Pavlo, Chair
David G. Andersen
Geoffrey J. Gordon
Michael Cafarella, University of Michigan

Thesis Summary: www.cs.cmu.edu/~dvanaken/papers/proposal.pdf